The interaction between structure and meaning in sentence comprehension

Frank, Stefan; Hoeks, J.C.J.

Published in:
Proceedings of the Cognitive Science Society

Publication date:
2019

The interaction between structure and meaning in sentence comprehension: Recurrent neural networks and reading times

Stefan L. Frank (s.frank@let.ru.nl)
Centre for Language Studies, Radboud University
Erasmusplein 1, 6525 HT Nijmegen, The Netherlands

John C. J. Hoeks (j.c.j.hoeks@rug.nl)
Faculty of Arts, University of Groningen
Oude Kijk in 't Jatstraat 26, 9712 EK Groningen, The Netherlands

Abstract

Recurrent neural network (RNN) models of sentence processing have recently displayed a remarkable ability to learn aspects of structure comprehension, as evidenced by their ability to account for reading times on sentences with local syntactic ambiguities (i.e., garden-path effects). Here, we investigate if these models can also simulate the effect of semantic appropriateness of the ambiguity's readings. RNN-based estimates of surprisal of the disambiguating verb of sentences with an NP/S-coordination ambiguity (as in 'The wizard guards the king and the princess *protects* ...') show identical patters to human reading times on the same sentences: Surprisal is higher on ambiguous structures than on their disambiguated counterparts and this effect is weaker, but not absent, in cases of poor thematic fit between the verb and its potential object ('The teacher baked the cake and the baker *made* ...'). These results show that an RNN is able to simultaneously learn about structural and semantic relations between words and suggest that garden-path phenomena may be more closely related to word predictability than traditionally assumed.

Keywords: garden-path sentences; self-paced reading; reading time; thematic fit; recurrent neural network; LSTM; surprisal

Introduction

Garden-path phenomena, in which a local structural ambiguity results in comprehension difficulty upon disambiguation, have been studied extensively in psycholinguistics. Traditionally, the garden-path effect has been explained in terms of syntactic structure building: When the ambiguity is encountered, the parser chooses the structure that later turns out to be incorrect, triggering a process of syntactic reanalysis (e.g., Frazier & Rayner, 1982). Nowadays, this process is often expressed in probabilistic terms: The syntactic interpretation of the sentence-so-far takes the form of a probability distribution over (all) possible structures, and processing a word comes down to redistributing the probability mass in light of the incoming linguistic information. In case of a garden-path sentence, the incorrect reading of the ambiguity receives a (much) higher probability than the correct one, which means that a lot of probability mass needs to be redistributed upon encountering the disambiguating word (Brouwer, Fitz, & Hoeks, 2010; Hale, 2001; Levy, 2008). This corresponds to high cognitive processing load.

In the probabilistic account of sentence processing sketched above, the amount of update in the probability distribution due to processing a word can be shown to equal the word's *surprisal*, which has therefore been proposed as relevant measure of cognitive processing difficulty during incremental language comprehension (Hale, 2001; Levy, 2008). Indeed, word surprisal correlates with word reading time in general, as long as it is estimated by an accurate-enough probabilistic language model. The model's underlying architecture does not appear to matter much: It can be a probabilistic grammar (Boston, Hale, Patil, Kliegl, & Vasishth, 2008; Demberg & Keller, 2008), a recurrent neural network (Goodkind & Bicknell, 2018; Monsalve, Frank, & Vigliocco, 2012), or even a simple n-gram model (Frank, 2017; Smith & Levy, 2013). However, it stands to reason that surprisal must be estimated by a model that builds syntactic structure (like a probabilistic grammar does) if it is to account for the garden-path phenomenon. After all, the garden-path effect is (allegedly) caused by structural reanalysis. Hence, a model that does not engage in structure building should not be able to explain the effect.

Recent results from Long Short-Term Memory models (LSTM; Hochreiter & Schmidhuber, 1997) cast doubt on this assumption. An LSTM is a recurrent neural network in which the flow of activation is controlled by gates with learned weights, making it better at learning long-distance dependencies than Elman's (1990) well-known Simple Recurrent Network. LSTMs have shown remarkable capability to deal with long-term structure (Gulordava, Bojanowski, Grave, Linzen, & Baroni, 2018), including correct predictions of reading-time effects in garden-path sentences. Van Schijndel and Linzen (2018) had an LSTM estimate surprisal on the disambiguating verb phrase in sentences such as 'The employee understood [that] the contract *would be* ...' (NP/S ambiguity; critical word in italics) and 'Even though the girl phoned[,] the instructor *was* ...' (NP/Z ambiguity). They found higher surprisal in the locally ambiguous sentences than in their unambiguous counterparts.[1] Futrell, Wilcox, Morita, and Levy (2018) show that an LSTM model can account for the garden-path effect in sentence pairs such as 'The witness [that was] examined *by the lawyer*' (RR/MV ambiguity). Moreover, the model correctly predicts the absence of a garden-path effect when the subject noun is inanimate, as in 'The evidence [that

[1] Futrell et al. (2019), however, report that LSTMs predict a weaker NP/Z garden-path effect when the ambiguous region is longer, contrary to what has been observed in human readers (Tabor & Hutchins, 2004).

was] examined *by the lawyer*'. This suggests that the LSTM learned not only the relative frequencies of the different structures but also how these frequencies interact with a lexical semantic property.

The current study goes beyond this work by looking at a different local structural ambiguity and, more importantly, its interaction with the thematic fit between an action (i.e., verb) and its potential patient (i.e., syntactic object). That is, we investigate the sensitivity of LSTMs to a semantic *relation* as opposed to a single word's semantic property.

Garden paths and thematic fit

Sentence (1a) is structurally ambiguous when the third noun phrase ('the princess') in encountered: It can be understood as part of the larger NP 'the king and the princess' or as the beginning of a new sentence clause. This is known as the NP/S-coordination ambiguity. The upcoming verb ('protects') disambiguates towards the S-coordination structure, which causes comprehension difficulty compared to the unambiguous variant (1b). In other words, (1a) is a garden-path sentence because readers initially prefer the NP-coordination reading (Frazier, 1987).

(1a) The wizard guards the king and the princess *protects* the prince with her life.

(1b) The wizard guards the king, and the princess *protects* the prince with her life.

Sentence pairs (2a) and (2b) are structurally identical to (1a) and (1b) but differ in an important respect: The NP-coordination reading, in which the teacher bakes both the cake and the baker, is semantically anomalous: Bakers are not usually baked objects. Would such poor thematic fit lead to an immediate S-coordination interpretation and, consequently, remove any comprehension difficulty in (2a) compared to (2b)?

(2a) The teacher baked the cake and the baker *made* twelve breads for the coming holidays.

(2b) The teacher baked the cake, and the baker *made* twelve breads for the coming holidays.

In an eye-tracking experiment, Hoeks, Hendriks, Vonk, Brown, and Hagoort (2006) investigated the processing of sentences with NP/S coordination ambiguities in Dutch, which is structurally identical to English in this respect. They found the expected garden-path effect in the Good Fit condition: Reading times were longer on sentences such as (1a) than on (1b). When thematic fit was poor (sentence pair 2a/b) the picture was less clear, but the authors concluded that there is also a garden-path effect in this condition, albeit weaker than that for the sentences with good thematic fit.

However, the reliability of this result is questionable because the garden-path effect on Poor Fit sentences never reached statistical significance on any of the investigated reading time measures; it was at best marginally significant for total reading time. Hoeks et al.'s conclusion was based on the presence of a main effect of Ambiguity (i.e., whether or not the sentence had a comma) in combination with the *absence* of a significant interaction with Thematic Fit. Hence, the claim that the garden-path effect also occurred in the Poor Fit sentences is in fact based on accepting the null hypothesis that there is no interaction.

The current study

We trained LSTM models on Dutch text corpora after which they estimated surprisal of the critical words in the experimental sentences of the Hoeks et al. (2006) study. In addition, we analysed unpublished self-paced reading data on these same sentences. Bayesian mixed-effects regression analyses revealed similar patterns for the surprisal values and reading times (RTs): They are larger in the locally ambiguous than unambiguous sentences and this difference is smaller (but not zero) in case of poor thematic fit than for sentence with good thematic fit. These findings demonstrate that poor thematic fit indeed reduces, but not completely removes, the garden-path effect caused by the NP/S-coordination ambiguity; and that these effects can be explained by the statistical word-order patterns that recurrent neural networks are able to learn from text corpora.[2]

Method

Self-paced reading experiment

Stimuli The stimulus set was identical to that of Hoeks et al. (2006). It consisted of 120 experimental sentences with a local NP/S coordination ambiguity. In 60 of the 120 sentences, the two nouns of the (potential) NP coordination were animates, making them semantically plausible objects of the verb. These were the Good Thematic Fit sentences (Example 1a, translated from Dutch). In the 60 Poor Fit sentences, in contrast, the verb had a strong selectional preference for an inanimate object and only the first noun of the potential NP coordination was inanimate (see 2a). Items were not matched between the Good Fit and Poor Fit conditions.

The sentence's critical word was the second verb (italicized in Examples 1 and 2), which always disambiguated towards the S-coordination reading. Unambiguous versions of the sentences were constructed by simply introducing a comma after the second noun (Examples 1b and 2b).

In addition to the experimental sentences, there were 200 filler sentences, 80 of which had unambiguous conjoined object NPs. In half of these fillers sentences, both object nouns were animate; in the other half the first object noun was inanimate and the second one animate, mimicking the order of inanimate/animate nouns in the Poor Fit condition. The other 120 fillers contained relative clauses.

Forty items were paired with a simple comprehension question in the form of a statement about the sentence. These were intended to ensure participants would read for comprehension.

[2]The LSTM models, self-paced reading data, and analysis code are available from `https://osf.io/npzc7`.

Participants One hundred and three native Dutch speaking undergraduate students from Radboud University participated in the experiment. The data from seven participants were excluded from analysis because they answered more than 20% of the comprehension questions incorrectly.[3] This left 96 participants with analysed data.

Procedure Each participant read 120 experimental sentences, 30 in each of the 2×2 (Ambiguity × Thematic Fit) conditions. Stimuli were presented using word-by-word, non-cumulative, moving window self-paced reading. The sentence appeared when the participant pressed a button, but only the first word was visible initially. All other characters (including the comma but excluding spaces and the end-of-sentence period) were replaced by hyphens. On each subsequent button press, the next word would be revealed and the previous word changed back to hyphens. If, after completing the sentence, a comprehension question appeared, the participant had to indicate by button press whether or not the statement was correct.

Neural network models

Training corpus Training sentences were selected from the NLCOW2014 corpus (Schäfer, 2015) which contains individual Dutch sentences crawled from the web. It is divided into seven slices with approximately 37 million sentences each. NLCOW14 treats punctuation marks as individual tokens, meaning that they are separated from the preceding and the following word. Because this is incorrect in case of the apostrophe, we preprocessed the corpus, reattaching apostrophes to the word to which they belong.[4]

For each slice, we extracted the 20,000 most frequent words without distinguishing between upper- and lower-case and ignoring any string containing a non-letter other than the hyphen or apostrophe. Next, this frequent-word list was joined with the set of word types in the Hoeks et al. (2006) stimuli. We then selected only and all corpus sentences that contain only words from the combined word list.[5] These sentences form the training data from that slice. The seven training sets comprised between 8.57 and 9.00 million sentences (108 to 115 million tokens) each.

Model architecture We trained one LSTM network on each of the seven training data sets for two epochs. All networks had a 300-dimensional input embedding layer, a 600-unit recurrent layer, a 300-unit non-recurrent layer between the recurrent and output layers, and softmax output layer with one unit for each word type in the training set. No attempt was made to optimize this architecture. The seven networks differed only in their output layer sizes and random initial connection weights.

After processing the first $t-1$ words of a sentence, the network's output activation for word unit w is its estimate of $P(w_t|w_{1\ldots t-1})$: the probability that word w will occur at position t given the word sequence (sentence context) w_1 to w_{t-1}. The surprisal of the actually occurring next word is defined as the negative logarithm of its occurrence probability: $\text{surprisal}(w_t) = -\log P(w_t|w_{1\ldots t-1})$.

Test sentences All seven networks estimated surprisal on all experimental sentences in both the Ambiguous (comma absent) and Unambiguous (comma present) condition. However, in spite of the training sentence selection method described above, 22 of the 120 experimental stimuli sentences contained one or more words not present in all seven training data sets. We replaced these words by semantically congruent words from the same syntactic category that did occur in all training sets. For example, in *De politie traceerde de dief* ('The police traced the thief') the verb *traceerde* was changed to *achtervolgde* ('chased').

Data analysis

We analysed the effect of Ambiguity on surprisal and RT by fitting Bayesian mixed-effects regression models using the R package `brms` (Bürkner, 2018). A positive regression coefficient for Ambiguity (i.e., $\beta_{\text{ambiguity}} > 0$) indicates higher surprisal or RT on Ambiguous than Unambiguous sentences, that is, a (predicted) garden-path effect.

The prior for $\beta_{\text{ambiguity}}$ was an improper flat distribution over the real numbers, as is the default in `brms`. For the RT analysis, it would have been justified to have the prior be informed by the Hoeks et al. (2006) results. However, we opted for a flat prior so that exactly the same analysis could be run for surprisal as for RT. The dependent variable was normalized so the intercept of the regression line is guaranteed to be 0. Hence, we set the strong prior of $\mathcal{N}(0, 0.1)$ over the intercept. We chose the Exponentially modified Gaussian family because of the positive skew in the dependent variables' distributions. The regression model included as random effects by-network and by-item random intercepts and random slopes of Ambiguity. Random-effect priors were the `brms` defaults.

Separate analyses were run for the Good and Poor Fit conditions, in addition to analyses including the factors Ambiguity, Fit, and their interaction. Both the Ambiguity and Fit factors were effect coded (± 0.5) with positive values for the Ambiguous and Good Fit conditions. Priors on the Fit and interaction coefficients were the default improper flat distribution.

Because self-paced reading often leads to so-called spillover effects, where comprehension difficulty on a word results in reading slowdown at a later word, we analysed RT on both the critical word and the immediately following word.

[3] There were in fact two versions of the experiment, which differed only in whether or not the comprehension questions were presented. Fifty-five of the the 103 participants took part in the version that included the questions. The data from the two experiment versions are combined in our analysis.

[4] In Dutch orthography, apostrophes can occur in the plural suffix *-'s* and in unstressed forms of pronouns (e.g., *m'n*, 'my') and determiners (e.g., *'n*, 'a').

[5] Single-word sentences and sentences containing over 50 words were excluded, as were sentences containing a punctuation token other than the period, comma, exclamation point, and question mark.

For completeness, we did the same for the surprisal analysis even though there is no reason why surprisal effects would spill over to the next word.

RTs below 50ms or over 4000ms were considered outliers and removed from analysis, but there were only four such data points (three on the critical word, one on the post-critical word).

Results

Effects of ambiguity and thematic fit

The two upper panels of Figure 1 show the posterior probability densities for the effect of Ambiguity on RT, in the Good and Poor Fit conditions. The reading time pattern is consistent with the conclusions Hoeks et al. (2006) draw from eye-tracking data on the same items: The ambiguity leads to a garden-path effect that is stronger in the Good than Poor Fit condition. The latter is apparent from the fact that, in Poor Fit sentences, the effect of Ambiguity occurs only on the critical word whereas it spills over to (and is even stronger on) the following word of Good Fit sentences. Table 1 presents the probability that there is indeed a garden-path effect in each of the Thematic Fit conditions, as well as the probability of an interaction such that the Ambiguity effect is larger in the Good Fit than Poor Fit condition.

This RT pattern is correctly predicted by the LSTM, as can be seen in the lower panels of Figure 1 as well as in Table 1. There is a clear effect of Ambiguity on surprisal in both the Good and Poor Fit conditions, and the evidence for an interaction between Ambiguity and Fit is very strong. Surprisal effects appear on the critical word rather than the post-critical word, which supports the claim that the post-critical RT effects are due to spillover of comprehension difficulty that arises at the critical word.

Effect of network training

As shown in Figure 2, it takes about 1 to 3 million training sentences for the garden-path effect and its interaction with thematic fit to appear. These effects continue to grow in size with additional training.

Table 1: Posterior probabilities of positive coefficients (i.e., $P(\beta > 0)$) of Ambiguity and its interaction with Thematic Fit.

			Word position	
Coefficient	Fit	Dep. Var.	Critical	Post-crit.
$\beta_{\text{ambiguity}}$	Good	RT	.98	$> .99$
		surprisal	$> .99$	.18
	Poor	RT	.93	.69
		surprisal	$> .99$	.32
$\beta_{\text{ambiguity}\times\text{fit}}$		RT	.78	$> .99$
		surprisal	$> .99$	.36

Item-level analysis

To investigate whether LSTM surprisal accounts for garden-path effects at the item level, surprisal was averaged per sentence over the seven fully trained networks, and log-transformed RTs were averaged per sentence over participants as well as over the critical and post-critical words. Figure 3 shows a scatter plot of average surprisal against average log-RT, excluding the 22 sentences that were adapted for LSTM processing. Clearly, the surprisal estimates are unable to explain garden-path effects at the level of individual sentences.

Discussion

Surprisal and reading time

Patterns of surprisal on the critical word matched the self-paced-reading results (as well as Hoeks et al.'s, 2006, eye-tracking data) albeit not at the individual item level. When comparing between experimental conditions, surprisal was higher when the sentence contained a local ambiguity (e.g., the LSTMs predict a garden-path effect) and this effect of Ambiguity was reduced (but not absent) when poor thematic fit between the verb and a following noun made the correct S-coordination reading more semantically appropriate before the disambiguating word. These results again demonstrate the power of RNNs to learn fairly subtle structural and semantic aspects of language, and thereby account for human processing behaviour.

The absence of effects on surprisal at the post-critical word supports the interpretation that the effect on RT here is caused by spillover from the critical word, as Hoeks et al. (2006) also conclude on the basis of their eye-tracking data. In that study, the authors found the garden-path effect to be more short-lived on Poor compared to Good Thematic Fit sentences. Our analysis of self-paced RTs shows the same pattern, in that the effect has disappeared on the post-critical word in the Poor Fit but not in the Good Fit condition. This suggests there may be a qualitative difference in the garden-path effects between the Thematic Fit conditions, that is not captured by the unidimensional surprisal measure.

Structural processing in RNNs

As explained in the Introduction, garden-path effects have been explained in terms of syntactic reanalysis, or probabilistically in terms of the redistribution of probability mass over syntactic structures. However, RNNs do not encode syntactic structure, at least not explicitly, so why did our networks correctly predict the garden-path effect?

One possibility is that the correspondence between surprisal and reading time is just an artefact of the experimental items. Possibly, the mere presence of a comma speeds up reading at the critical word, but less so in the Poor Fit Sentences, without any relation to the garden-path phenomenon. However, even if this is the case, it leaves unexplained why at least three other garden-path effects have been explained by

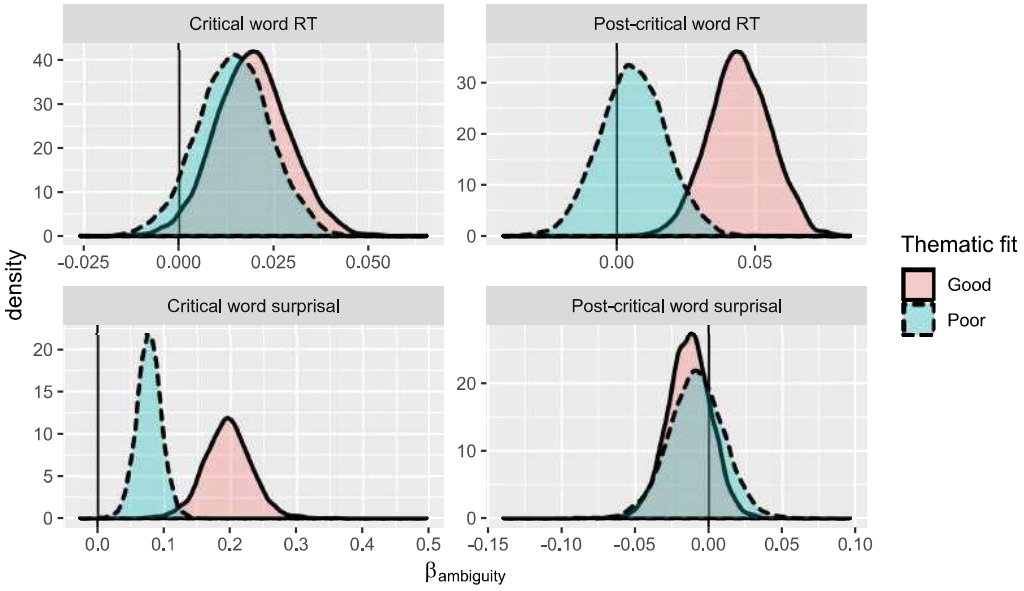

Figure 1: Posterior probability densities of the Ambiguity coefficient. Top: effect on RT; bottom: effect on word surprisal. Left: effect at critical word; right: effect at post-critical word.

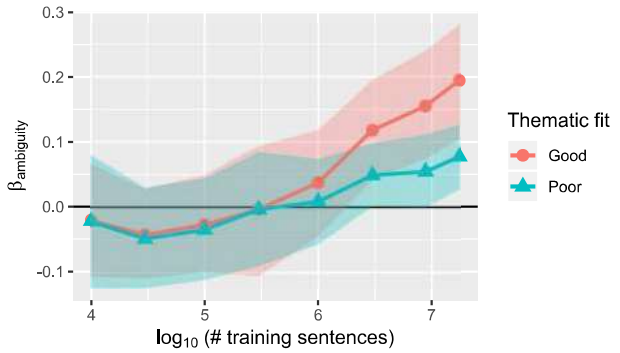

Figure 2: Estimated coefficient of Ambiguity at the critical verb in the surprisal analysis, as a function of number of training sentences and thematic fit. Shaded areas represent 95% Credible Intervals.

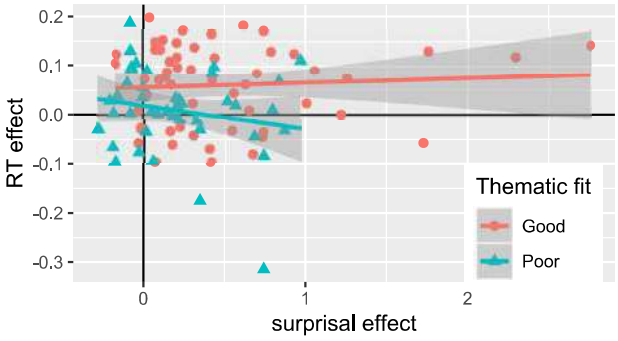

Figure 3: Garden-path effects in surprisal estimates and log-transformed RT, with regression line per Thematic Fit condition.

LSTM surprisal (Futrell et al., 2018; Van Schijndel & Linzen, 2018).

Alternatively, garden-path effects could be merely due to incorrect next-word prediction, as reflected in high surprisal on the disambiguating word. This would imply that there is no qualitative difference between comprehension difficulty due to a garden-path and due to an unlikely word co-occurrence. However, this seems implausible considering that ERP studies have shown that garden-pathing leads to a P600 effect (Osterhout & Holcomb, 1992; Osterhout, Holcomb, & Swinney, 1994) while higher surprisal in non-garden-path sentences corresponds to a stronger N400 component (Delaney-Busch, Morgan, Lau, & Kuperberg, 2019; Frank, Otten, Galli, & Vigliocco, 2015). Moreover, the initially preferred, but incorrect, reading of the ambiguity in a garden-path sentence can 'linger' (Christianson, Hollingworth, Halliwell, & Ferreira, 2001; Patson, Darowski, Moon, & Ferreira, 2009) which shows that such an interpretation was indeed entertained.

Possibly, being led up the garden path *also* results in incorrect next-word prediction and the *reading time effect* that comes with garden pathing actually reflects the resulting surprisal increase rather than the update of a structure or interpretation. However, this is not a particularly satisfying explanation as it would mean that the cognitive work of reanalysis

is itself not reflected in longer reading time.

Hence, we tentatively conclude that the LSTMs learn representations that capture relevant aspects of sentence structures/interpretations. As words come in, the network performs probabilistic, incremental reinterpretation, and generates word surprisal values that reflect the amount of representation update required to incorporate the word into the sentence representation under construction.

Conclusion

Word surprisal values estimated by LSTM models mirrored human reading times on garden-path sentences, predicting both the garden-path effect itself and its interaction with the manipulation of thematic fit between a verb and its potential object noun. This finding yet again demonstrates LSTMs' ability to extract structural aspects of language by learning to do next-word prediction in flat, unannotated text. Investigations of the neural networks' internal state are needed to substantiate this claim. If such an investigation fails to reveal evidence of structure representations in the networks, this would raise doubt about the necessity for structure building and revision in an explanation of garden-path phenomena.

Acknowledgments

The work presented here was funded by the Netherlands Organisation for Scientific Research (NWO) Gravitation Grant 024.001.006 awarded to the Language in Interaction Consortium.

References

Boston, M. F., Hale, J., Patil, U., Kliegl, R., & Vasishth, S. (2008). Parsing costs as predictors of reading difficulty: An evaluation using the Potsdam Sentence Corpus. *Journal of Eye Movement Research*, *2*, 1–12.

Brouwer, H., Fitz, H., & Hoeks, J. (2010). Modeling the noun phrase versus sentence coordination ambiguity in Dutch: Evidence from surprisal theory. In *Proceedings of the 2010 Workshop on Cognitive Modeling and Computational Linguistics* (pp. 72–80). Uppsala, Sweden: Association for Computational Linguistics.

Bürkner, P.-C. (2018). Advanced Bayesian multilevel modeling with the R package brms. *The R Journal*, *10*, 395–411.

Christianson, K., Hollingworth, A., Halliwell, J. F., & Ferreira, F. (2001). Thematic roles assigned along the garden path linger. *Cognitive Psychology*, *42*, 368–407.

Delaney-Busch, N., Morgan, E., Lau, E., & Kuperberg, G. R. (2019). Neural evidence for Bayesian trial-by-trial adaptation on the N400 during semantic priming. *Cognition*, *187*, 10–20.

Demberg, V., & Keller, F. (2008). Data from eye-tracking corpora as evidence for theories of syntactic processing complexity. *Cognition*, *109*, 193–210.

Elman, J. L. (1990). Finding structure in time. *Cognitive Science*, *14*, 179–211.

Frank, S. L. (2017). Word embedding distance does not predict word reading time. In *Proceedings of the 39th Annual Conference of the Cognitive Science Society* (pp. 385–390). Austin, TX: Cognitive Science Society.

Frank, S. L., Otten, L. J., Galli, G., & Vigliocco, G. (2015). The ERP response to the amount of information conveyed by words in sentences. *Brain and Language*, *140*, 1–11.

Frazier, L. (1987). Syntactic processing: evidence from Dutch. *Natural Language and Linguistic Theory*, *5*, 519–559.

Frazier, L., & Rayner, K. (1982). Making and correcting errors during sentence comprehension: Eye movements in the analysis of structurally ambiguous sentences. *Cognitive Psychology*, *14*, 178–210.

Futrell, R., Wilcox, E., Morita, T., & Levy, R. (2018). RNNs as psycholinguistic subjects: Syntactic state and grammatical dependency. *arXiv preprint*. Retrieved from `http://arxiv.org/abs/1809.01329`

Futrell, R., Wilcox, E., Morita, T., Qian, P., Ballesteros, M., & Levy, R. (2019). Neural language models as psycholinguistic subjects: Representations of syntactic state. In *Proceedings of the 18th Annual Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies.*

Goodkind, A., & Bicknell, K. (2018). Predictive power of word surprisal for reading times is a linear function of language model quality. In *Proceedings of the 8th Workshop on Cognitive Modeling and Computational Linguistics* (pp. 10–18). Salt Lake City, UT: Association for Computational Linguistics.

Gulordava, K., Bojanowski, P., Grave, E., Linzen, T., & Baroni, M. (2018). Colorless green recurrent networks dream hierarchically. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies* (pp. 1195–1205). New Orleans, LA: Association for Computational Linguistics.

Hale, J. T. (2001). A probabilistic Early parser as a psycholinguistic model. In *Proceedings of the Second Conference of the North American Chapter of the Association for Computational Linguistics* (Vol. 2, pp. 159–166). Pittsburgh, PA: Association for Computational Linguistics.

Hochreiter, S., & Schmidhuber, J. (1997). Long short-term memory. *Neural Computation*, *9*, 1735–1780.

Hoeks, J., Hendriks, P., Vonk, W., Brown, C., & Hagoort, P. (2006). Processing the noun phrase versus sentence coordination ambiguity: Thematic information does not completely eliminate processing difficulty. *The Quarterly Journal of Experimental Psychology*, *59*, 1581–1599.

Levy, R. (2008). Expectation-based syntactic comprehension. *Cognition*, *106*, 1126–1177.

Monsalve, I. F., Frank, S. L., & Vigliocco, G. (2012). Lexical surprisal as a general predictor of reading time. In *Proceedings of the 13th Conference of the European Chapter of the Association for Computational Linguistics* (pp. 398–408). Avignon, France.

Osterhout, L., & Holcomb, P. J. (1992). Event-related brain

potentials elicited by syntactic anomaly. *Journal of Memory and Language*, *31*, 785–806.

Osterhout, L., Holcomb, P. J., & Swinney, D. A. (1994). Brain potentials elicited by garden-path sentences: Evidence of the application of verb information during parsing. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *20*, 786–803.

Patson, N., Darowski, E., Moon, N., & Ferreira, F. (2009). Lingering misinterpretations in garden-path sentences: evidence from a paraphrasing task. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *35*, 280–285.

Schäfer, R. (2015). Processing and querying large web corpora with the COW14 architecture. In P. Bański, H. Biber, E. Breiteneder, M. Kupietz, H. Lüngen, & A. Witt (Eds.), *Proceedings of the 3rd Workshop on the Challenges in the Management of Large Corpora* (pp. 28–34). Mannheim, Germany: Institut für Deutsche Sprache.

Smith, N. J., & Levy, R. (2013). The effect of word predictability on reading time is logarithmic. *Cognition*, *128*, 302–319.

Tabor, W., & Hutchins, S. (2004). Evidence for self-organized sentence processing: Digging-in effects. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *30*, 431–450.

Van Schijndel, M., & Linzen, T. (2018). Modeling garden path effects without explicit hierarchical syntax. In T. Rogers, M. Rau, X. Zhu, & C. W. Kalish (Eds.), *Proceedings of the 40th Annual Conference of the Cognitive Science Society* (pp. 2603–2608). Austin, TX: Cognitive Science Society.